10-301/601: Introduction to Machine Learning Lecture 22: Hidden Markov Models

Henry Chai

7/18/23

Front Matter

- Announcements
  - PA5 released 7/13, due 7/20 at 11:59 PM
- Recommended Readings
  - Murphy, Chapters 17.1 - 17.5

Structured Data

- For many machine learning tasks, the training data will have some implicit structure or ordering.
  - Time series data
  - Text data
  - Audio/video data

- $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right)\right\}_{n=1}^{N}$ where each training data point consists of multiple observations in *sequence*:

$$\boldsymbol{x}^{(n)} = \left[\boldsymbol{x}_1^{(n)}, \dots, \boldsymbol{x}_{T_n}^{(n)}\right]$$

$$\boldsymbol{y}^{(n)} = \left[y_1^{(n)}, \dots, y_{T_n}^{(n)}\right]$$

Part-of-Speech (PoS) Tagging

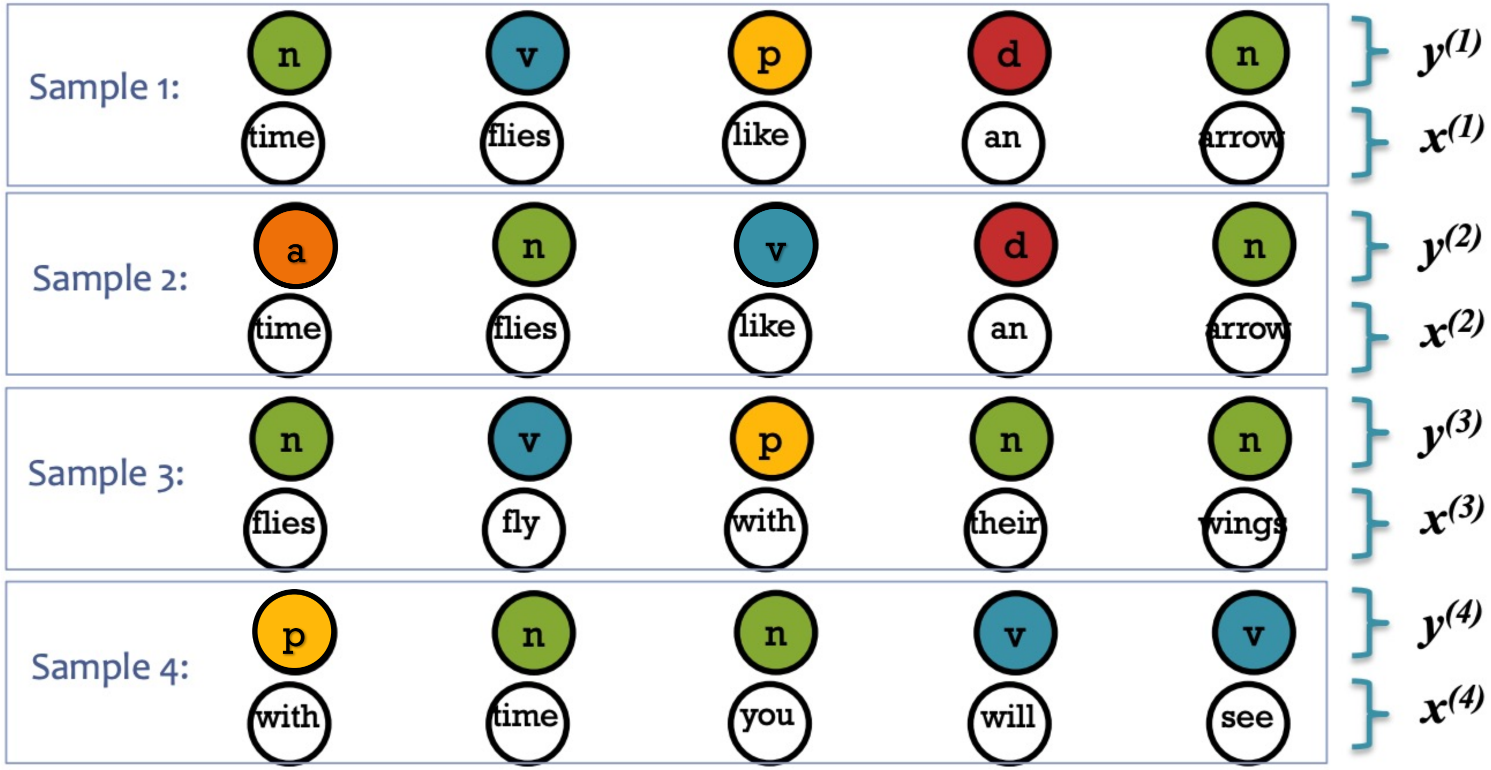
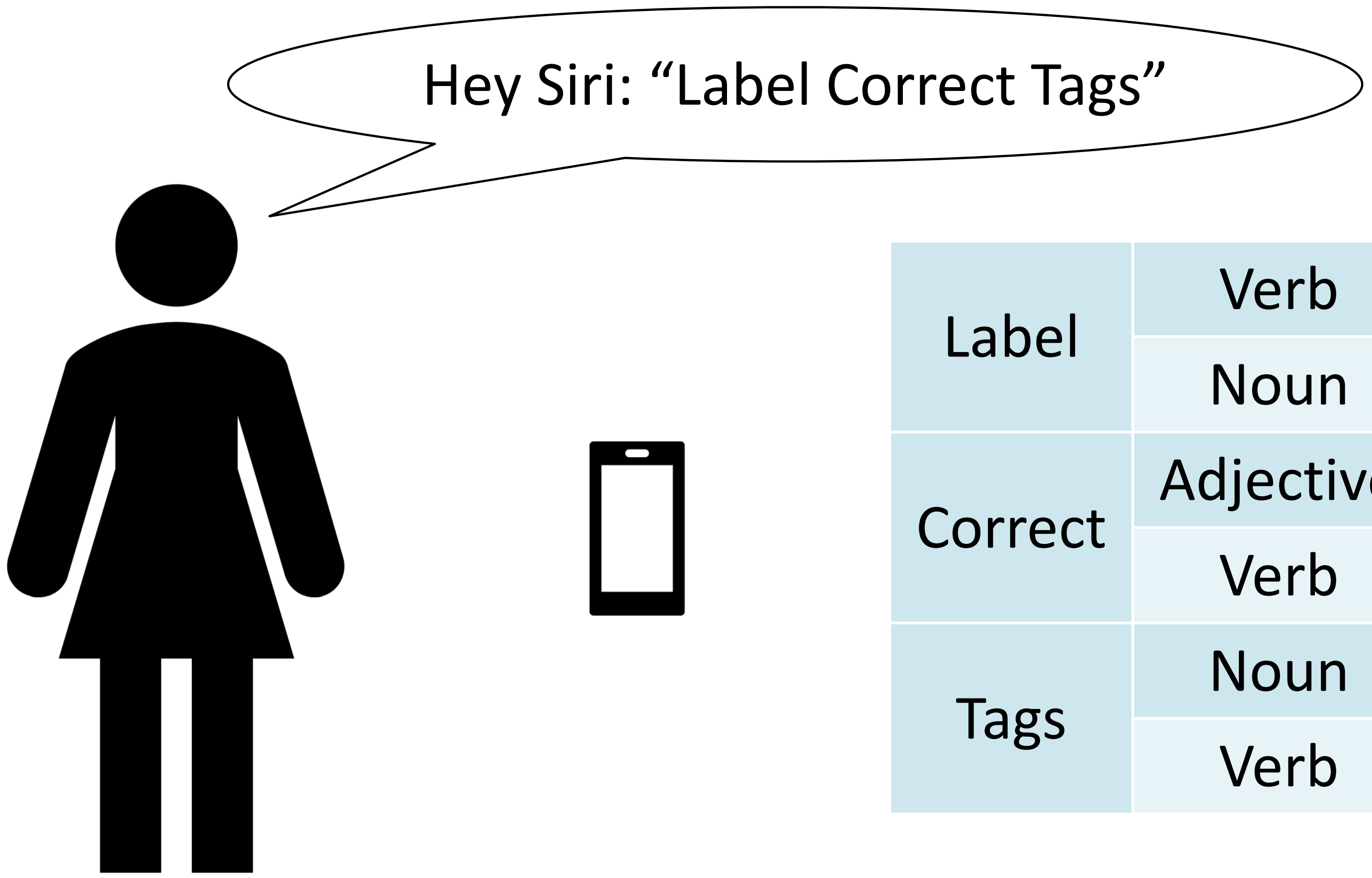


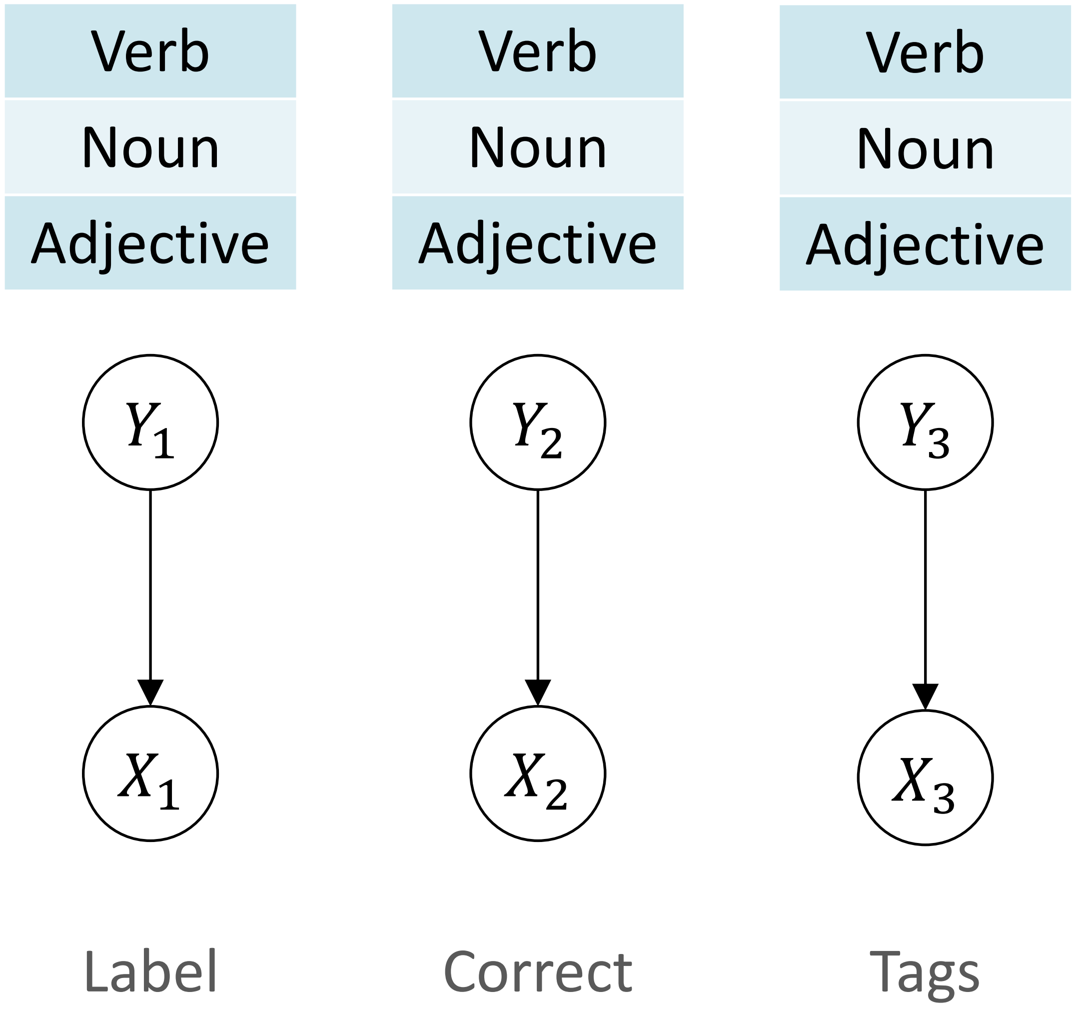
Figure courtesy of Matt Gormley

Part-of-Speech (PoS) Tagging: Example

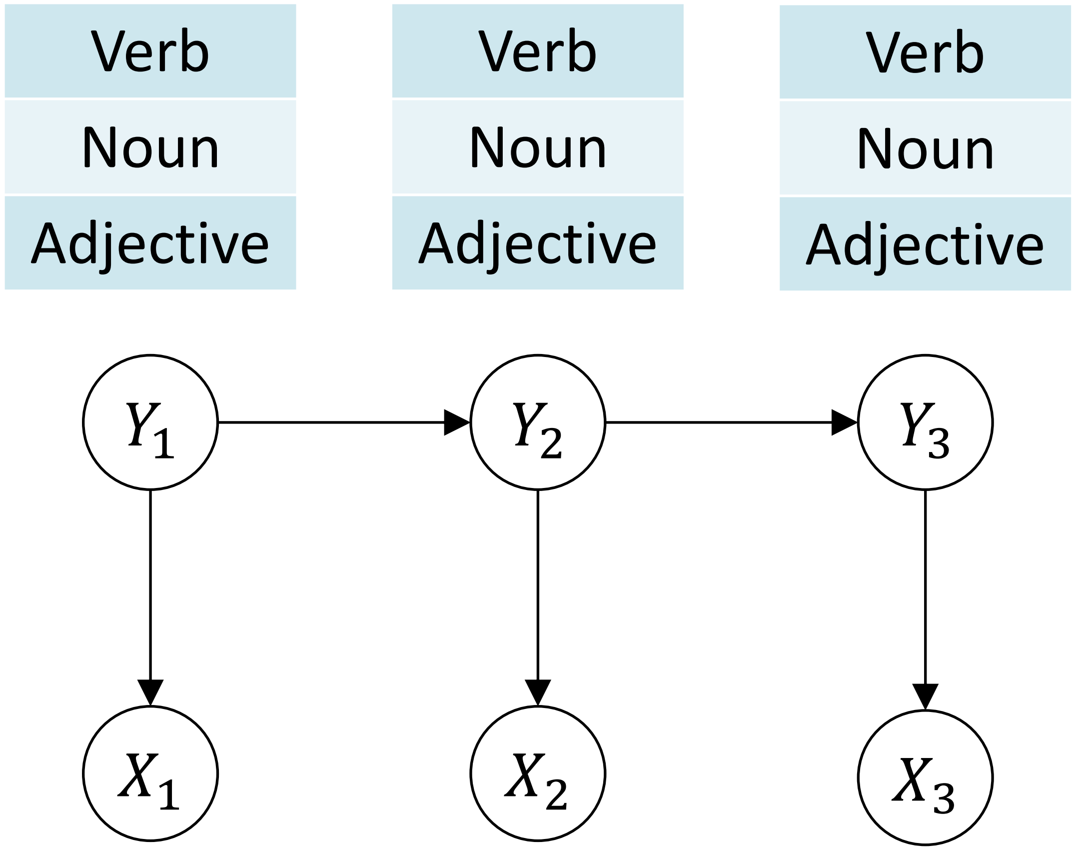


Label	Verb
	Noun
Correct	Adjective
	Verb
Tags	Noun
	Verb

Naïve Bayes for PoS Tagging

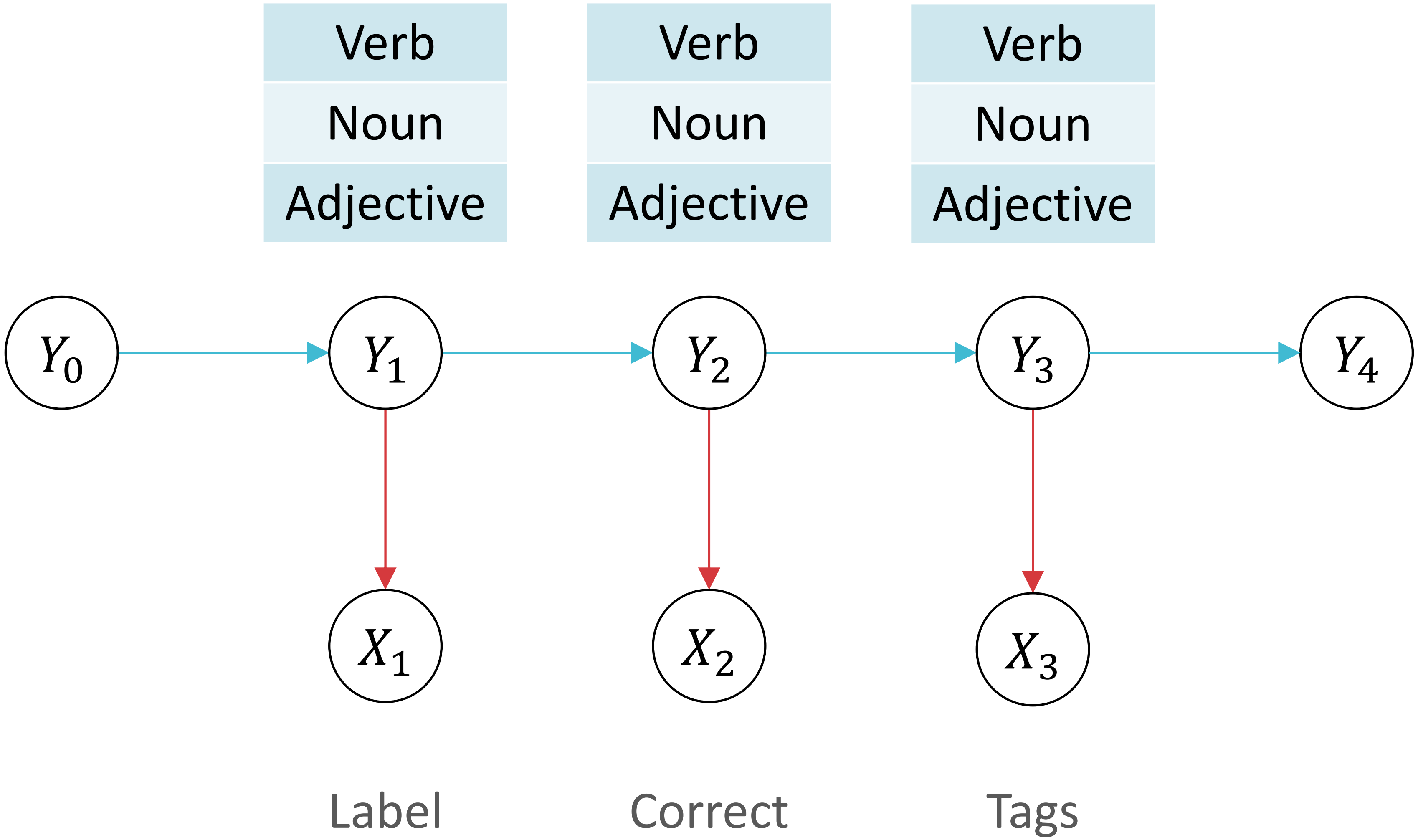


(Dynamic) Bayesian Network for PoS Tagging



Label Correct Tags

Hidden Markov Models for PoS Tagging



Hidden Markov Models

- Two types of variables: observations (observed) and states (hidden or latent)
  - Set of states usually pre-specified via domain expertise/prior knowledge: $\{s_1, \dots, s_M\}$
  - Emission model:
    - Current observation is conditionally independent of all other variables given the current state: $P(X_t | Y_t)$
  - Transition model (Markov assumption):
    - Current state is conditionally independent of all earlier states given the previous state (Markov assumption): $P(Y_t | Y_{t-1}, \dots, Y_0) = P(Y_t | Y_{t-1})$

Hidden Markov Models vs. Bayesian Networks

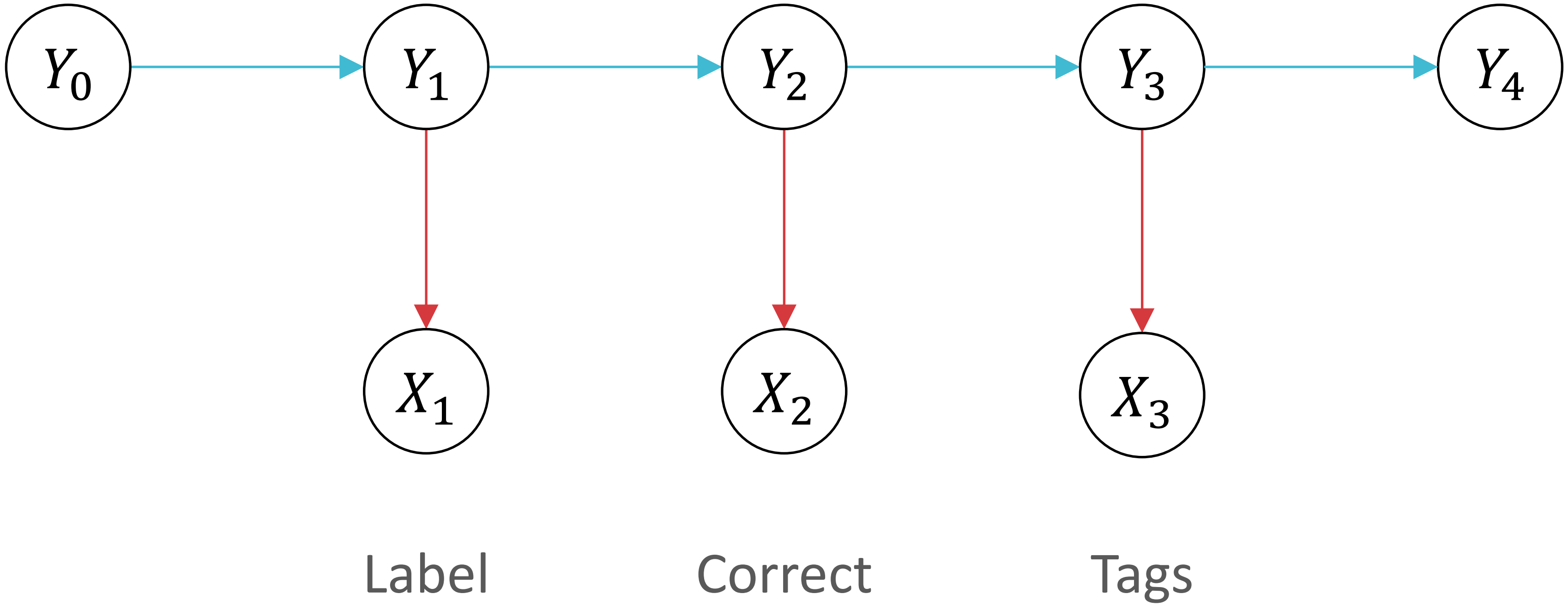
- Two types of variables: observations (observed) and states (hidden or latent)
  - Set of states usually pre-specified via domain expertise/prior knowledge: $\{s_1, \dots, s_M\}$
  - Emission & transition models are fixed over time steps

$$P(X_t | Y_t = s_j) = P(X_{t'} | Y_{t'} = s_j) \; \forall \, t, t'$$

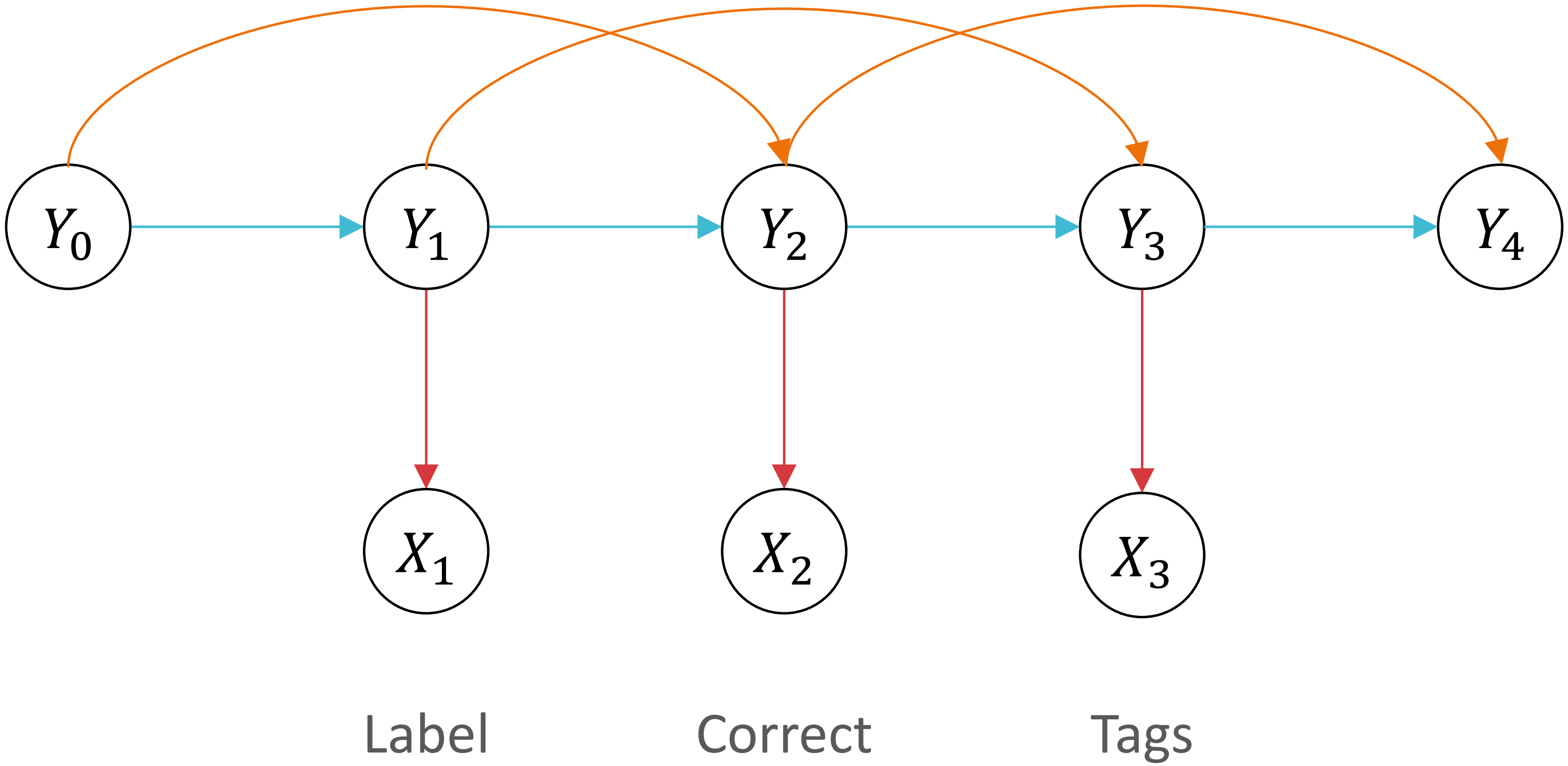
$$P(Y_t | Y_{t-1} = s_j) = P(Y_{t'} | Y_{t'-1} = s_j) \; \forall \, t, t'$$

    - Parameter reuse makes learning efficient
    - Can handle sequences of varying lengths

1^{st} Order Hidden Markov Models for PoS Tagging



2nd Order Hidden Markov Models for PoS Tagging



Hidden Markov Models: Outline

- How can we learn the conditional probabilities used by a hidden Markov model?
- What inference questions can we answer with a hidden Markov model? (tomorrow)
  1. Computing the distribution of a single state (or a sequence of states) given a sequence of observations
  2. Finding the most-probable sequence of states given a sequence of observations
  3. Computing the probability of a sequence of observations

Learning the Parameters (Fully-observed)

- Given C possible observations and M possible states plus special START/END states, how many parameters do we need to learn?

Given C possible observations and M possible states plus special START/END states, how many parameters are in the emission matrix, A?

- MC
- $M(C-1)$
- C^2
- $C(C-1)$

Given C possible observations and M possible states plus special START/END states, how many parameters are in the transition matrix, B?

- M^2
- $M(M-1)$
- $M(M+1)$
- $(M+1)^2$

Learning the Parameters (Fully-observed)

- Given C possible observations and M possible states plus special START/END states, how many parameters do we need to learn?

	s_1	$\cdots$	s_M
o_1	a_{11}	$\cdots$	a_{1M}
o_2	a_{21}	$\cdots$	a_{2M}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_C	a_{C1}	$\cdots$	a_{CM}

Emission matrix, A

$$a_{ij} = P(X_t = o_i | Y_t = s_j)$$

	START	s_1	$\cdots$	s_M
s_1	b_{10}	b_{11}	$\cdots$	b_{1M}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_M	b_{M0}	b_{M1}	$\cdots$	b_{MM}
END	$b_{(M+1)0}$	$b_{(M+1)1}$	$\cdots$	$b_{(M+1)M}$

Transition matrix, B

$$b_{ij} = P(Y_t = s_i | Y_{t-1} = s_j)$$

Learning the Parameters (Fully-observed)

- $\mathcal{D} = \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})\}_{n=1}^{N}$
- Set the parameters via MLE

	s_1	$\cdots$	s_M
o_1	a_{11}	$\cdots$	a_{1M}
o_2	a_{21}	$\cdots$	a_{2M}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_C	a_{C1}	$\cdots$	a_{CM}

Emission matrix, A

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} N_{X_t=o_i,\, Y_t=s_j}}{\sum_{t=1}^{T} N_{Y_t=s_j}}$$

	START	s_1	$\cdots$	s_M
s_1	b_{10}	b_{11}	$\cdots$	b_{1M}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_M	b_{M0}	b_{M1}	$\cdots$	b_{MM}
END	$b_{(M+1)0}$	$b_{(M+1)1}$	$\cdots$	$b_{(M+1)M}$

Transition matrix, B

$$\hat{b}_{ij} = \frac{\sum_{t=1}^{T+1} N_{Y_t=s_i, Y_{t-1}=s_j}}{\sum_{t=1}^{T+1} N_{Y_{t-1}=s_j}}$$

Key Takeaways

- HMMs are an instantiation of (dynamic) Bayesian networks where certain parameters are shared
  - Parameters can be set by MLE